VocSegMRI: 多模态学习在实时 MRI 中的精确声道分割

刘戴琪 *
Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany

Tomás Arias-Vergara
Pattern Recognition Lab, FAU
GITA Lab, Universidad de Antioquia UdeA
Medellín, Colombia

Johannes Enk
Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany

Fangxu Xing
Harvard Medical School / MGH
Boston, USA

Maureen Stone
Department of Orthodontics and Pediatrics
University of Maryland School of Dentistry, USA

Jerry L. Prince
Department of Electrical and Computer Engineering
Johns Hopkins University
Baltimore, USA

Jana Hutter
Smart Imaging Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany

Andreas Maier
Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany

Jonghye Woo
Department of Orthodontics and Pediatrics
University of Maryland School of Dentistry, USA

Paula Andrea Pérez-Toro*
Pattern Recognition Lab, FAU
GITA Lab, Universidad de Antioquia UdeA
Erlangen, Germany

ABSTRACT

准确地在实时磁共振成像（rtMRI）中分割发音结构仍然具有挑战性，因为大多数现有方法几乎完全依赖于视觉线索。然而，同步的声学和音韵信号提供了补充上下文，可以丰富视觉信息并提高精度。在这篇论文中，我们介绍了 VocSegMRI，这是一个多模态框架，通过交叉注意力融合整合视频、音频和音韵输入以进行动态特征对齐。为了进一步增强跨模态表示，我们在推理时即使没有可用的音频模式也纳入了一个对比学习目标来提升分割性能。在 USC-75 rtMRI 数据集的一个子集上评估了我们的方法，达到了最先进的性能，Dice 得分为 0.95，并且 95 百分位豪斯多夫距离（HD_{95}）为 4.20 毫米，优于单模态和多模态基线。消融研究表明交叉注意力和对比学习对分割精度和鲁棒性的贡献。这些结果突显了集成多模态建模在精确声门分析中的价值。

*Corresponding authors: daiqi.deutschfau.liu@fau.de, Preprint submitted to ICASSP

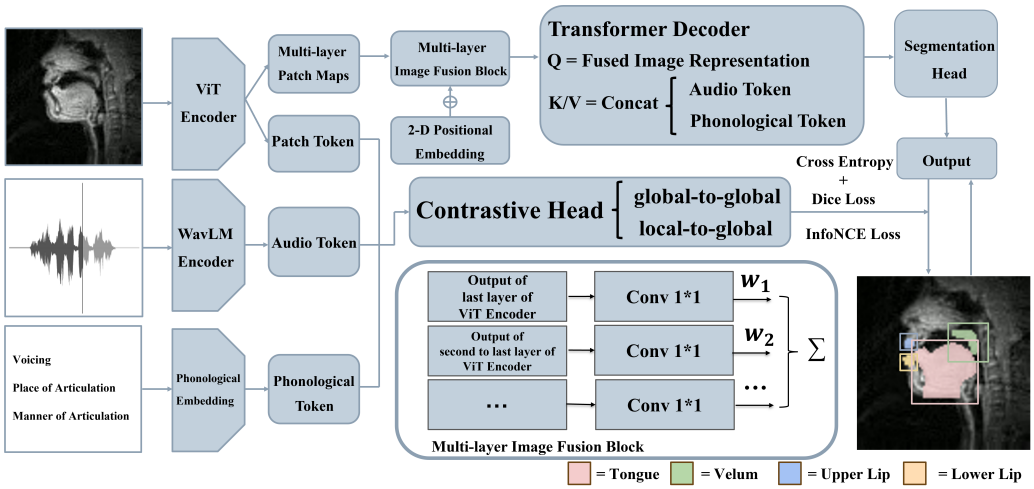


图 1: 所提出的 VocSegMRI 模型概述，包含交叉注意力融合和对比监督。

Keywords 分

割，多模态学习，实时 MRI，声道

1 介绍

实时磁共振成像（rtMRI）在语音研究中越来越被采用，提供了一种非侵入性的、高时间分辨率的整个发声道在连续言语过程中的视图 [1] 。准确分割发声道结构不仅对音位学和语言学研究至关重要，而且对于临床应用也非常重要，如舌切除术前的手术规划以及监测帕金森病患者的发音器官退化情况 [2–4] 。除了成像之外，声学信号和音系类别特征还捕捉到了结构化的发音特性，例如发音部位和方式，提供了可以与 rtMRI 结合进行更精确分析的符号上下文 [5, 6] 。

早期提取发音轮廓的努力主要依赖于手动或半自动边界追踪。典型的工作流程包括在参考帧中手工标注空气-组织边界，然后通过非线性优化将这些轮廓传播到后续的帧上 [7] 。其他方法则是通过检查叠加在 MRI 图像上的预定义网格线上像素强度来分配标签，或使用主动形状模型捕捉预期的解剖轮廓 [8] 。这个过程不仅耗时且劳动密集型，而且容易出错，需要大量的人工监督。最近，提出了利用深度学习进行分割的方法 [9–12] 。特别是，[9] 使用全卷积网络（FCNs）来标注空气-组织边界，而 [9] 还将边界像素分配给特定的发音器官。其他作品，如 [10] , 分别利用了类似于原始 FCN [13] 和在 [14] 中变体的 FCNs 来标注空气-组织界面。Matthieu Ruthven 等开发了一种基于 U-Net 的分割模型，识别出六个发声道结构，实现了 0.85 的 Dice 系数 [11] 。最近，一种结合了 rtMRI 与同步语音音频的多模态系统被引入，该系统使用了一个全卷积架构并结合 Transformer 为基础的融合方法，在说话人无关的发声道分割上设立了新的基准 [12] 。

在这项工作中，我们提出了 VocSegMRI，一个用于 rtMRI 中发音器分割的多模态框架，它集成了视觉、声学和音系输入。我们的贡献有三点：(i) 一个融合了交叉注意力机制的框架，使视觉编码器能够专注于音频和音素流中的互补信息；(ii) 全局和局部应用的双层对比学习目标，以提高跨模态对齐和一致性；以及 (iii) 在 USC-75 数据集上的系统评估，在该数据集中，我们的模型实现了 0.95 的 Dice 系数和 4.20 毫米的 HD_{95}，超过了强大的单模态和多模态基线。这些结果证明了综合多模态建模对于更精确地进行声道分割的有效性。

2 材料与方法

2.1 数据

我们利用了来自 USC-75 数据集的五名参与者（一名男性和四名女性）的数据进行训练和领域内评估。每位参与者被指示阅读*彩虹*和*北风和太阳*段落 [15, 16]。实时 MRI 数据以 **83.28 帧/秒**获得，平面分辨率为 2.4 毫米（84×84 像素），切片厚度为 6 毫米 [17]。成像使用了一台 **1.5 T**GE Signa Excite 扫描仪；同步音频通过光纤麦克风以 **20 千赫兹**的频率录制，硬件锁定到扫描仪时钟并进行了降噪处理。二值分割掩模是手动标注的，并由一位在该领域具有公认专长的语言治疗专家进一步审查。音素是从音系分类器的输出中获得的，与音频流中的相应音素标签对齐，并通过人工校正进一步细化 [5, 6]。

为了扩大训练集，对每张图像应用了几何和强度变换相结合的数据增强。增强后的图像随后被调整为 224×224 像素的分辨率。每个数据集都基于其最小和最大强度值进行了单独归一化处理，以确保像素强度在 [0,1] 范围内。结果，共获得了 **14,406** 张可用于训练和评估的有效图像。

2.2 VocSegMRI

我们提出的模型的整体框架如图 1 所示。它在一个跨注意力 Transformer 架构中整合了视觉、声学和音系信息，并辅以对比学习。在编码器阶段，基于`google/vit-base-patch16-224-in21k` 检查点的预训练 Vision Transformer (ViT) 从 rtMRI 帧 [18, 19] 中提取空间表示，同时同步音频使用基于`microsoft/wavlm-base-plus` 检查点的预训练 WavLM 模型编码 [20]，并利用轻量级 MLP 映射音系特征。后两者结合并投影形成多模态记忆标记，这些标记通过 Transformer 解码器中的跨注意力层与图像查询交互 [21]，从而实现感知模态的融合以进行精细分割。为了进一步加强跨模态对齐，对比模块将图像、音频和音系标记投影到共享潜在空间中。这种监督鼓励模型利用声学-音系指导学习更具判别性的视觉特征。因此，该模型支持仅使用视频进行推理，而无需额外的模态。最终训练目标集成了交叉熵、Dice 和对比损失。源代码将在接受后发布。

3 实验和结果

在五个可用的参与者中，我们采用了留一说话人策略。为了确保公平比较，所有实验均在相同的预处理管道下进行，并且没有进行任何后处理。在训练过程中，音频编码器被冻结以保留其预训练的声学表示，而最初被冻结的 ViT 编码器则逐步解冻，以便逐渐微调以适应 rtMRI 领域。所有模型使用 AdamW 进行优化，学习率为 $1e-4$，批量大小为 16，在 NVIDIA RTX A100 GPU 上训练。在验证集上训练直至收敛，并应用提前停止（耐心=15）以防止过拟合。

我们设计了三个实验组，结果分别呈现在表 1 的上、中、下部分。评估指标包括交并比（IoU）、Dice 系数、平均对称表面距离（ASSD）和 HD_{95}。所有指标均以均值 ± 标准差（std）的形式报告。在**仅图像基线**中，我们比较了几种最先进的（SOTA）分割模型的性能，这些模型仅使用逐帧 rtMRI 图像作为输入 [22–27]。其中，ViT-base 和 nnU-Net 获得了最高的 IoU 为 **0.86**，其次是 ResNet（0.84）和 SAM-Med2D（0.83）。在**连接融合**设置中，我们通过连接每种模式各自的编码器输出来实现多模态融合。我们使用不同的输入模式组合评估了分割性能。添加音频或音系类别特征都使得 IoU 比 ViT（0.86）有所提升，提升幅度有所不同。最佳性能是通过使用所有三种模态实现的（IoU=0.89）。最后，在**消融研究**配置中，我们单独评估了每个组件的贡献。**交叉**配置仅包含交叉注意力，而**对比**仅引入对比学习。我们提出的模型**语音分割 MRI** 结合了这两种机制，实现了整体最佳性能。获得最高的 Dice 分数（**0.95**）和最低的 ASSD（**1.52**）及 HD_{95}（**4.26**）。

图 2 展示了提出的 VocSegMRI 模型在测试集上的每类分割性能。如图 2(a) 所示，该模型在*舌头*和*帆状结构*类上实现了最高的 Dice 分数，中位数值分别超过 0.95 和 0.93，并且四分位距较窄，表明分割性能稳定。相比之下，*上唇*和*下唇*表现出更大的变异性并且通常 Dice 分数较低，中位数分别约为 0.85 和 0.83。在图 2(b) 中也观察到了类似的趋势，*下唇*类的中位数 HD_{95} 超过 4 毫米，分布范围广泛且存在许多异常值。

表 1: 不同模型和输入模式的分割性能。ASSD/HD 以毫米（mm）为单位给出。

模型	交并比 ↑	骰子 ↑	ASSD↓	HD_{95}↓
		仅图像基线		
U-Net	0.81 ± 0.06	0.89 ± 0.03	3.64 ± 0.77	8.08 ± 1.21
Swin UNETR	0.82 ± 0.05	0.89 ± 0.02	3.21 ± 0.58	7.54 ± 1.17
SAM-Med2D	0.83 ± 0.05	0.91 ± 0.02	3.23 ± 0.61	7.67 ± 1.12
ResNet-50	0.84 ± 0.04	0.91 ± 0.02	3.04 ± 0.57	7.22 ± 1.08
nnU-Net	0.86 ± 0.03	0.93 ± 0.02	1.74 ± 0.43	5.37 ± 1.08
ViT-base	0.86 ± 0.02	0.94 ± 0.01	2.11 ± 0.58	4.73 ± 0.91
		连接融合		
V	0.86 ± 0.02	0.94 ± 0.01	2.11 ± 0.58	4.73 ± 0.91
VA	0.87 ± 0.03	0.91 ± 0.03	2.40 ± 0.49	5.81 ± 0.98
VP	0.86 ± 0.04	0.89 ± 0.02	3.21 ± 0.58	7.54 ± 1.17
VAP	0.89 ± 0.02	0.94 ± 0.01	2.19 ± 0.41	5.00 ± 0.97
		消融研究		
Cross-Att	0.90 ± 0.02	0.95 ± 0.01	1.83 ± 0.37	4.91 ± 1.00
Contrastive	0.89 ± 0.02	0.93 ± 0.01	2.03 ± 0.44	4.35 ± 0.87
VocSegMRI	**0.91 ± 0.01**	**0.95 ± 0.01**	**1.52 ± 0.31**	**4.26 ± 0.88**

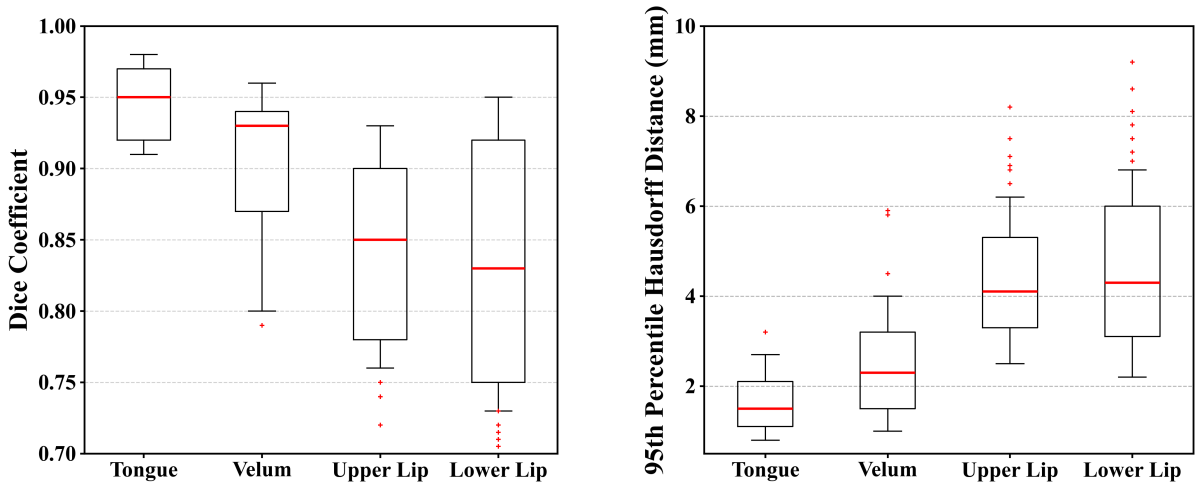
V: 视频。A: 音频。P: 音系学。Cross-Att: 交叉注意力。

我们选择了四个代表性的模型进行比较。图 3 展示了*舌头*和下唇的定性分割结果，突出了假阳性（FP）和假阴性（FN）区域，并报告了精确率/召回率值，而 nnU-Net 在所有模型中表现最差。具体来说，对于*舌头*，nnU-Net 达到了 0.91 的精确率和 0.80 的召回率，而对于下唇，精确率为 0.42，召回率为 0.97。将所有三种模式作为输入导致模型性能提升。然而，VocSegMRI 在这两种结构上始终实现了最佳权衡，整体精确率为 **0.85**，召回率为 **0.98**。

4 讨论与结论

实验结果表明，我们提出的多模态分割框架**语音分割 MRI** 有效提升了分割性能，并且与已建立的基线相比表现出了优越的分割精度。仅使用逐帧 rtMRI 图像作为输入时，编码器的选择影响了性能，ViT 在单模态模型中实现了最高的 IoU。这验证了基于 Transformer 的视觉特征提取对于解剖结构的有效性。

在**连接融合**设置中，我们调查了纳入额外模态的影响。将音频或音系特征与视觉信息拼接起来，在仅视频基线（IoU 0.86）上有所改进，当所有模态同时使用时效果最佳（IoU 0.89）。值得注意的是，单独添加音系特征



a b

图 2: 所提出的 VocSegMRI 模型在测试集上的每类分割性能。(a) Dice 系数。(b) 第 95 百分位 Hausdorff 距离（HD_{95}）。

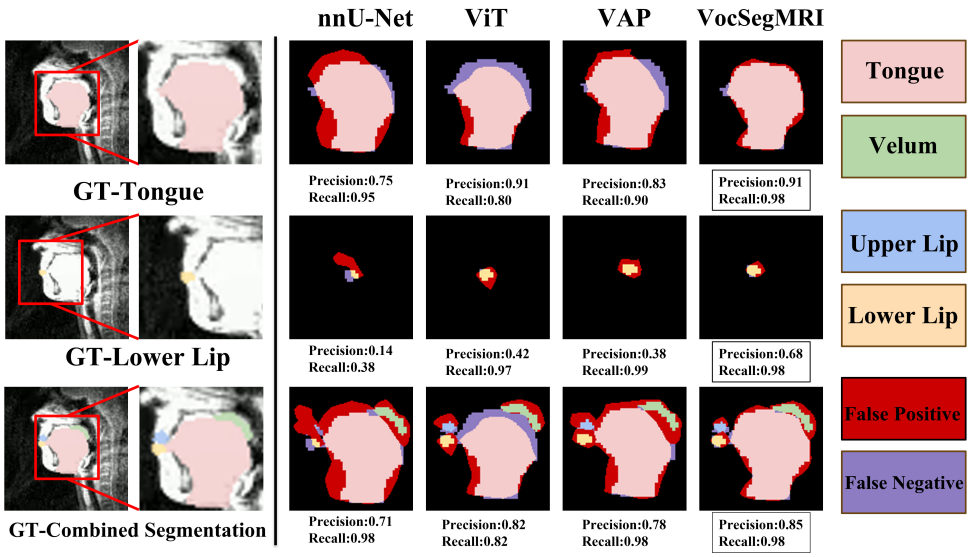


图 3: 地面真相（GT）和预测分割的定性比较之一 rtMRI 帧

相比整合音频信号所带来的性能提升较小。这种有限的改进可能归因于学习音系标签与视觉发音区域之间有意义映射的难度，这降低了它们引导模型注意力到特定区域的能力。总体而言，这些结果证实了多模态输入提供了互补信息，通过提供动态发音线索和结构化的音系信息增强了分割质量。

消融研究探索了先进的融合机制和对比监督。用交叉注意力替换简单的拼接，使图像查询能够有选择地整合来自音频和音素流的互补信息，同时增加对比学习目标进一步对齐跨模式的表示。结果，**VocSegMRI** 达到了最佳的整体性能，Dice 分数为 **0.95**，ASSD 为 **1.52 毫米**，HD_{95} 为 **4.26 毫米**，超过了单模态和简单的拼接基线。值得注意的是，模型在所有度量上也表现出较低的标准偏差，表明这些增益在不同说话者中是一致的，并突显了框架的稳健性。

类级分析（图 2）表明，分割对于较大的结构如舌头和帆状结构更为准确，而较小的发音器官（上/下唇）由于在图像中像素表示非常低以及视觉线索较弱，仍然具有挑战性。例如，在图 3 所示的例子中，尽管召回值较高（分别为 0.38 和 0.97），下唇使用 nnU-Net 仅达到 0.14 的精度，而基于 ViT 的基线则达到 0.42 的精度，这反映了大量误报。相比之下，舌头在不同模型中表现出较高的精度（0.75 – 0.91）和高召回率。定性比较进一步说明，单模态模型容易出现明显的误报和漏报，特别是在解剖边界处。虽然简单的拼接部分缓解了这些错误，但结合对比学习的交叉注意力使分割更加精确和均衡，将下唇的精度提高到 0.68，并且整体的精度召回率折衷提高到 0.85/0.98，从而大大减少了所有结构中的假阳性和假阴性。

总结来说，我们提出了一种新颖的实时磁共振成像多模态分割框架，通过交叉注意力融合和对比监督整合了视觉、声学和音系信息。在 USC-75 数据集上的实验展示了最先进的性能，证实了模态感知对齐的有效性。对比组件确保即使语音信息退化或缺失（如舌切除患者）时也能实现可靠的分割。尽管取得了这些进展，小的低对比度结构的准确分割仍然是一个挑战。未来工作将探索自适应注意力机制、时间建模、针对性数据增强和领域泛化策略，以提高对代表性不足发音器和未见说话者的性能，为稳健且独立于说话人的声道分析铺平道路。

参考文献

[1] Asterios Toutios et al., "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, pp. e6, 2016.

[2] Adam C Lammert et al., "Investigation of speed-accuracy tradeoffs in speech production using real-time magnetic resonance imaging.," in *Interspeech*, 2016, pp. 460–464.

[3] Christina Hagedorn et al., "Characterizing post-glossectomy speech using real-time mri," in *International Seminar on Speech Production, Cologne, Germany*, 2014, pp. 170–173.

[4] Catherine P Browman et al., "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[5] Daiqi Liu et al., "Audio–vision contrastive learning for phonological class recognition," in *International Conference on Text, Speech, and Dialogue*. Springer, 2025, pp. 60–71.

[6] Tomás Arias-Vergara et al., "Contrastive learning approach for assessment of phonological precision in patients with tongue cancer using mri data," 2024, p. 927.

[7] Vikram Ramanarayanan et al., "Analysis of speech production real-time mri," *Computer Speech & Language*, vol. 52, pp. 1–22, 2018.

[8] Mathieu Labrunie et al., "Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning," *Speech Communication*, vol. 99, pp. 27–46, 2018.

[9] Krishna Somandepalli et al., "Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images.," in *Interspeech*, 2017, pp. 631–635.

[10] Renuka Mannem et al., "Air-tissue boundary segmentation in real time magnetic resonance imaging video using a convolutional encoder-decoder network," in *ICASSP*. IEEE, 2019, pp. 5941–5945.

[11] Matthieu Ruthven et al., "Deep-learning-based segmentation of the vocal tract and articulators in real-time magnetic resonance images of speech," *Computer Methods and Programs in Biomedicine*, vol. 198, pp. 105814, 2021.

[12] Rishi Jain et al., "Multimodal segmentation for vocal tract modeling," *arXiv preprint arXiv:2406.15754*, 2024.

[13] Yiheng Zhang et al., "Fully convolutional adaptation networks for semantic segmentation," in *Proceedings of the IEEE CVPR*, 2018, pp. 6810–6818.

[14] Jimei Yang et al., "Object contour detection with a fully convolutional encoder-decoder network," in *Proceedings of the IEEE CVPR*, 2016, pp. 193–202.

[15] John Garofolo et al., "Darpa timit acoustic-phonetic continuous speech corpus cd-rom TIMIT," 1993.

[16] Frederic L Darley et al., *Motor speech disorders*, Saunders, 1975.

[17] Yongwan Lim et al., "A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images," *Scientific data*, vol. 8, no. 1, pp. 187, 2021.

[18] Bichen Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," 2020.

[19] Jia Deng et al., "Imagenet: A large-scale hierarchical image database," in *2009 IEEE CVPR*. IEEE, 2009, pp. 248–255.

[20] Sanyuan Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[21] Ashish Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Fabian Isensee et al., "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[23] Kaiming He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR*, 2016, pp. 770–778.

[24] Alexander Kirillov et al., "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[25] J Cheng et al., "Sam-med2d. arxiv 2023," *arXiv preprint arXiv:2308.16184*.

[26] Ali Hatamizadeh et al., "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 272–284.

[27] Olaf Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.